Research Paper
*Impact Factor - **2.262***
Peer Reviewed Journal

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

AN INNOVATIVE APPROACH IN SEARCHING OF RESEARCH PAPERS USING TEXT CLUSTERING WITH FEATURE SELECTION

B.S.Sangeetha* K.N.Nithya* N.Suganya Devi* A.Shyamala Gowri*
**Assistant Professor, Shri Sakthikailassh Women's College, Salem,(TN) India.*

Abstract
While accessing a list of research papers from a variety of wide spread journals, the title entered by the users and the most highlighted words specified along with the title are considered to search multifarious documents distributed worldwide. To deploy this process, a number of Text Mining techniques are applied. One among the technique is implementation of Clustering Algorithm that helps to search for the text documents very efficiently. The results are presented to the users based only on the similarity of the text documents present in the title and the key terms specified by the users. Our proposed research considered here is to highlight the inclusion of most recent papers based on the year of publication. This proposal is implemented by using K-Means algorithm. In this paper, we create a common cluster that moves all the similar text documents according to the year of publication.

Key Words: Text Clustering, Feature Selection, k-Means Clustering, Improved K-Means Clustering

INTRODUCTION

Due to the enormous growth of information World Wide Web plays a very vital role to provide exact and efficient information to the millions of people accessing the Internet. Therefore for the people to search research papers, to reach their point of understanding, after searching across several topics, it takes a long duration of time. Given the high number of researches and the increasing amount of information on the web, it becomes very important to organize this large amount of information into meaningful clusters referenced by distinct categories [6]. Hence we can use concept of Web Mining integrated with Text Clustering to get rid of this problem. The paper is divided in to seven sections.

Section I explains the significance of Text Clustering and its various functions. Section II explains the use of Feature Selection. Section III elaborates on the use of K-Means Clustering. Section IV is our proposed work that signifies the use of improved k-Means clustering to arrange the text documents according to the year of publication.

SECTION I
TEXT CLUSTERING

The term Data Mining refers to a set of algorithms that extract meaningful information based on the user requirements. Normally information is available in raw data form which has to be refined, filtered and structured in order to generate patterns that are understandable and relevant to the users. Data Mining is most often associated with the broader process of Knowledge Discovery in Databases(KDD), "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data".[1]. There are number of mining process available like classification, clustering. Here we are going to apply the Text Clustering mechanism.

Clustering is a division of data into groups of related objects. Every group is called cluster, which having objects that are similar between themselves and different to objects of further groups [2].Clustering is one of an efficient technique used in Document Clustering. Clustering helps to find the similarity of documents and assigns a common class to the data objects. It helps to generate patterns to huge number of datasets.By implementing an efficient algorithm, it is possible for the researchers to get in to their needs based on the topics rather than searching in all the areas which is meaningful.Thus Clustering is a useful technique for the discovery of data distribution and patterns in the underlying data sets [1].

For the clustering technique to work efficiently, we must have to choose a good search engine and a better database to access and retrieve best results. We have used MySQLs database server that helps to access all kinds of websites that allows the use of multithreading and multiuser environment.

SECTION II
FEATURE SELECTION

Feature selection or Attribute selection is applied in the situations when high-dimensionality of data is present that needs intense searching. High dimension of data increases the complexity of understanding the dataset itself and applying the algorithm, since many algorithms are sensitive to largeness or high-dimensionality or both[7].The data that are present may

be irrelevant or meaningless data. These types of datasets will not provide the exact requirement even when efficient algorithms are being applied. Therefore these redundant or raw data has to be removed before it is going to produce valid information to the users. For this purpose, we apply Feature Selection process. It is pre-processing step, where we filter the unrequited data and extract only the useful information for document clustering. Those data that fits this requirement is moved to k-Means Clustering process.

There are three major benefits of feature selection (FS): (1) improves the prediction performance of the predictors; (2) helps predictors do faster and more cost-effective prediction; and (3) provides a better understanding of the underlying process that generated data [2]. There are number of feature selection algorithms available like Correlation-Based Filter Approach, Clustering Ensemble for Unsupervised Feature Selection, Text Clustering with Feature Selection (**TCFS**) by Using Statistical Data, Novel Unsupervised Feature Selection method for Bioinformatics Data Sets through Feature Clustering Data, K-MEANS ALGORITHM [2]. Here we are going to make use of K-Means Clustering along with Feature Selection to assign weights for our proposed work.

SECTION III
k-MEANS CLUSTERING

K-Means Clustering is a partitioning technique that divides the dataset in to a group of clusters. This method is very helpful in finding the similarity between the documents. It is used to search for the class label that the given text belong and puts the text as an object under the class label. K-Means algorithm will be used here to send the parameters for classification of research papers [3]. Once the given text is identified it searches the class label that has a similar meaning to the text, it adds the text in the form of an object under the class label. This class label acts as an identifier to find the object text as and when needed. This algorithm forms a group of clusters and all the dataset that are identical to any of the cluster is placed sunder that cluster.

The algorithm assigns each point to the cluster whose center is nearer to it called as “centroid”. The centroid is the mean value of all the points of clusters present. All the clusters that are closest to the centroid are arranged. The process is repeated to find the next centroid value for the rearranged clusters. The clusters are again grouped on the new centroid value. The process is done repeatedly, till the centroid value is no more to be found. At every pass of the algorithm, each data value is assigned to the nearest partition based upon some similarity parameter such as Euclidean distance of intensity [5].
The algorithm steps are;

1. Choose the number of clusters, k.
2. Randomly generate k clusters and determine the cluster centers, or directly generate k random points as cluster centers.
3. Assign each point to the nearest cluster center.
4. Re-compute the new cluster centers.
5. Repeat steps 2 and 3 until convergence.

The centroids are calculated in k-means algorithm, arithmetic mean of the cluster all points of a cluster with the given distance measure distances are computed [3]. E.g. Euclidean distance. The distance function between two points x= (a1,b1) and y=(a2,b2) is defined as-

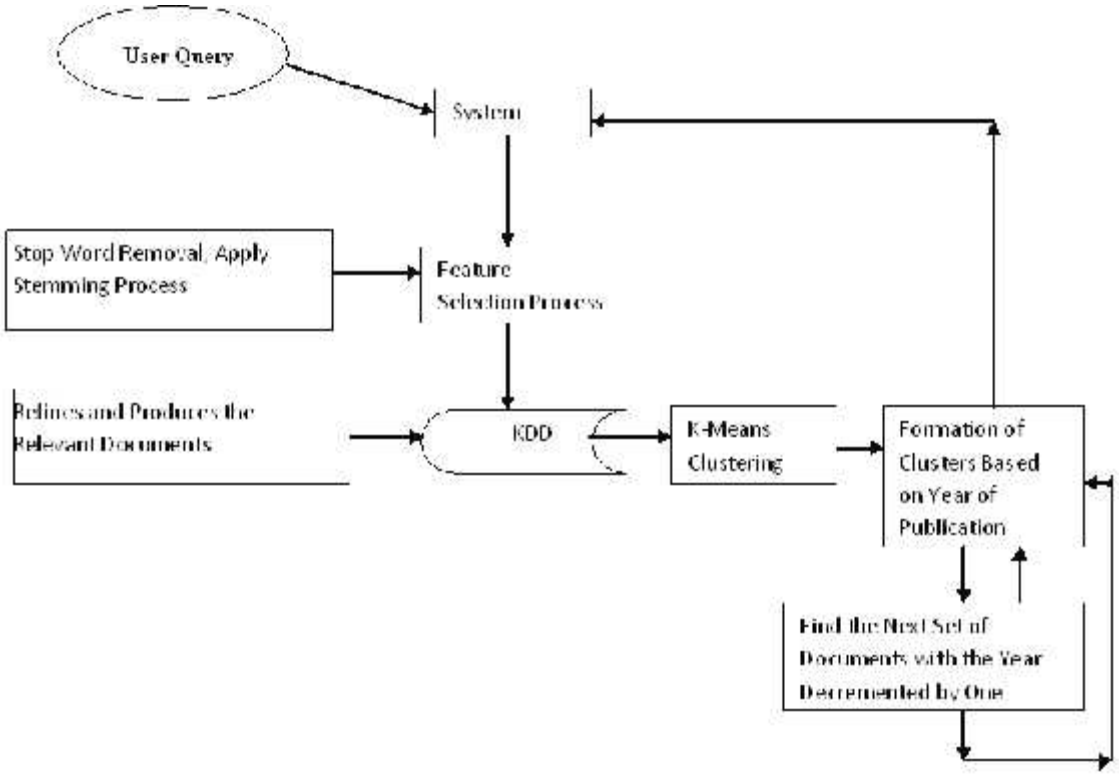
$$\mathbf{p(x,y)=|a2-a1|+|b2-b1|}$$

k-Means algorithm has the ability to generate large number of clusters with ease and simplicity. The algorithm is used to send the parameters for classification of research papers [3]. The results are also produced accurately even if large number of datasets when considered.

SECTION IV
PROPOSED WORK: IMPROVED k-MEANS ALGORITHM

The system architecture is designed to retrieve the user query regarding on a specified topic related to any area of research.

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877



The Knowledge Discovery DataBase[KDD], that stores enormous information is used to discover research papers based on the search patterns created from the user requirement in Feature Selection process. The weights are assigned to the terms based on Cosine values along with the additional information about the current year of publications are automatically added. This architecture shows how a cluster is generated based on the year of publications of the PDF documents. The procedure is as follows;

1. Retrieval of User Query.
2. The data entered is moved to the Preprocessing stage, which is a Feature Selection stage that involves removal of meaningless data and terms. In the preprocessing stage, apart from stop word elimination and stemming, a weight estimation function, that calculates the term weight and semantic weight are included. Term weight is estimated using TF/IDF (Term Frequency/Inverse Document Frequency] values that utilize information about term and number of times (n) it appears in the document [1].
3. The attributes extracted are moved to pattern recognition phase that extracts the various papers from the DataBase.
4. From the list of papers generated we apply the following clustering algorithm to generate a cluster based on year.

We have used an aggleromative algorithm. Each of the singleton cluster denotes the text documents with the measured similarity. The input datasets are

- Current Year
- Title
- Vital terms if any.

Implementation of Algorithm

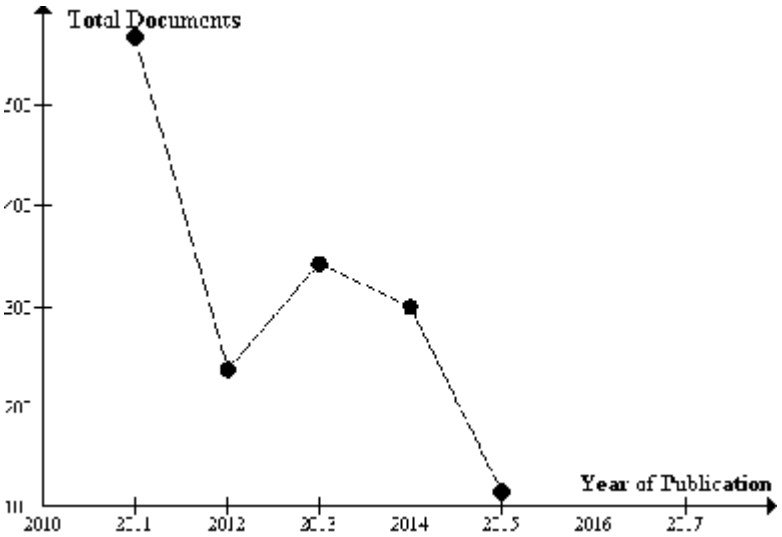
Input: D= {d1, d2, d3,...., dn} //set of 'n' data points
K //Number of needed clusters
Output: Set of 'k' clusters arranged
Steps:

1. Y = Current Year
2. d = Y added with data set.
3. N=5
4. For i = 1 to 5
5. For j = 1 to d

6. Compute Centroid 'C'= |Xi+1-Xi|+|Yi+1-Yi|
7. Move 'd' to the closest Centroid 'C'
8. Repeat the steps 3 to 5 until the clusters are formed
9. Y=Y-1
10. Goto 4
11. Stop

Under each cluster, we can get all the research papers that fall under a particular year category. The following is the graph that demonstrates arrangement of clusters based on the year of publication. The table shows the list of websites accessed for five consecutive years.

RESULT AND DISCUSSION



Graphical Representation

Table:1

Documents	Years
115	2015
300	2014
343	2013
237	2012
569	2011

Description of Document Sets

CONCLUSION

The proposed work will be very efficient and useful for the people who want an up-to-date information about the various research papers developed recently. Based on this orderly arrangement, the research people has the maximum possibility to get into newer and innovative ideas than becoming tiresome in viewing the older research papers at first sight. Our future work is to enhance the further readability by finding the total number of times a particular journal is visited by implementing the Web Page Ranking Algorithm.

REFERENCES

1. "Clustering Technique in Data Mining for Text Documents", Ms.J.Sathya Priya Ms.S.Priyadharshini, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (1) , 2012
2. "Efficiently Clustering of High Dimensional Data using Attribute Selection", Miss Palak, V. Desai,Mr Arpit Rana, *(IJCSITS), ISSN: 2249-9555,Vol. 2, No.6, December 2012*
3. "Searching Research Papers Using Clustering and Text Mining", Jadhav Bhushan G, Warke Pushkar U, Kuchekar Shivaji P, Kadam Nikhil V, IJETAE,Volume 4, Issue 4, April 2014
4. "A Comparative Study to Find a Suitable Method for Text Document Clustering",Mrs.S.C.Punitha and Dr.M.Punithavalli, IJCSIT Volume 3, No 6, Dec 2011
5. "Improved K-means Algorithm for Searching Research Papers", Sachin Shinde, Bharat Tidke, International Journal of Computer Science & Communication Networks, Volume 4(6),197-202
6. "A Centroid and Relationship based Clustering for Organizing Research Papers", Damien Hanyurwimfura, Liao Bo, Dennis Njagi and Jean Paul Dukuzumuremyi, International Journal of Multimedia and Ubiquitous Engineering Vol.9, No.3 (2014), ISSN: 1975-0080 IJMUE Copyright ⓒ 2014 SERSC
7. "Improved Clustering Approach Based on Fuzzy Feature Selection", IEEE 2007, Naiju Wu, Xiuyun Li, Jie Yang, Pengs Liu.